DOCUMENT RESUME

ED 388 680 TM 023 712

AUTHOR Benoit, Joyce; Yang, Hua
TITLE A Redefinition of Portfolio Assessment Based upon Purpose: Findings and Implications from a Large-Scale Program.
PUB DATE Apr 95
NOTE 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Accountability; Criteria; *Educational Assessment; Educational Policy; Educational Theories; Elementary Secondary Education; Models; *Portfolio Assessment; School Districts; *Scoring; Standardized Tests; Test Construction; Testing Programs; Test Reliability; *Test Use; Test Validity
IDENTIFIERS Dallas Independent School District TX; Education Consolidation Improvement Act Chapter 1; *Large Scale Programs; *Performance Based Evaluation

ABSTRACT

The use of portfolio assessment in the Dallas (Texas) Independent School District and the future of portfolio assessment are discussed. A literature review is followed by a description of the development process that preceded the Chapter 1 portfolio assessment of the Dallas schools. Portfolio results are then compared to the standardized measures available within the school district, and the issues of reliability and validity are discussed. The policy implications of portfolio use are also discussed. The experiences of the school district lead to the conclusion that the current theoretical model for portfolio assessments should be changed. When a portfolio is used for accountability purposes, it must be designed from the top-down with clearly defined criteria and appropriate rubrics. Reliability in portfolio assessment comes only from a well-structured and carefully scored portfolio. A portfolio designed to improve instruction and learning must be designed from the bottom up. It is not possible to use a single portfolio assessment system to accomplish both goals. (Contains 1 figure and 19 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

JOYCE BENOIT

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC).

ED 388 680

A Redefinition of Portfolio Assessment Based Upon Purpose: Findings and Implications from a Large-Scale Program

Joyce Benoit, Ph.D.
Hua Yang, Ph.D.

Dallas Independent School District
Department of Evaluation and Testing
3801 Herschel Avenue
Dallas, Texas 75219
(214) 522-8220
(214) 559-1980 (Fax)

Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, California, April 18-22, 1995

BEST COPY AVAILABLE

Nationwide, portfolio assessment as a form of performance-based assessment has been used experimentally in several states or school districts and is drawing more and more attention from researchers and policy-makers. Despite its promised merit of being able to link performance assessment with classroom instruction, studies on the effects of the pioneer portfolio programs revealed a number of concerns over its inherent weakness, especially when it was applied to large-scale evaluation (e.g., Koretz, Klein, McCaffrey, & Stecher, 1993). In general, the concerns concentrate on three related aspects:

1. What is a portfolio assessment? Too many times the term portfolio assessment is interpreted differently. A clear definition needs to be developed. Does it include student work samples? Best work samples? Is a standard piece included in all portfolio or is the work collected randomly for each student? What is the teacher's role? What is the student's role? What timeline should be used for data collection? Does a portfolio show growth on a few objectives or measure final performance on many objectives? Without a clear definition the portfolio outcomes cannot be clearly interpreted, assessed, or even defended against critics.

2. What is the major goal of portfolio assessment? Performance assessment or instructional improvement? Despite the theoretically conventional wisdom that "good assessment brings good instruction," it becomes a completely different story in actual practice. Can we meet these conflicting goals at the same time? It is clear that there must be some compromise between the inherently conflicting goals (assessment and instruction) which are assigned to portfolio assessment. What is acceptable to both sides? What will make them balanced? Is there any bridge connecting the two sides?

3. Whether or not portfolio assessment is able to provide high quality data about student performance? How to ensure a reasonable level of reliability in portfolio assessment when there is a serious lack of consistent understanding among teachers about the criteria and the standards applied in the evaluation? What measures of concurrent validity are available and appropriate?

It has been three years since portfolio assessment was implemented in the Dallas Independent School District (DISD) as an outcome measure in determining the effectiveness of the Chapter 1 program. As Chapter 1 program evaluators in the DISD, we felt the difficulties accompanied by a lack of knowledge as we first set up the design of portfolio assessment. We shared the frustrations of many of the Chapter 1

teachers when they implemented the project, trying to make sense from the confusion. However, step by step, as our experiences increased, the quality of the portfolio assessment in the DISD Chapter 1 program has been improved. In our work, we are particularly aware of the three common concerns raised by the studies on portfolio assessment. These are also the issues we have been dealing with in improving our work. These are the issues that will be discussed in this study.

This paper will first explore the literature for related issues on the topic of portfolio assessment followed by a description of the developmental process of the DISD Chapter 1 portfolio assessment. The portfolio results will then be compared to the standardized measures available within the District and the issues of reliability and validity will be discussed. Finally, after discussing the unresolved issues in the DISD portfolio assessment, we conclude this study with its policy implications. Based upon our in-depth experiences with portfolio assessment in the DISD, we will recommend a change to the current theoretical model for portfolio assessments. In our opinion, grass-roots development of a portfolio assessment will improve instruction, but will not serve the purposes of large-scale assessment. The policy trend in many states that mandates Districts use a top down approach to reach bottom up goals will only create a great deal of confusion about appropriate uses of portfolio assessment.

Literature Review Focused on the DISD Assessment Portfolio

Recently, there has been increased interest in using performance assessments that are similar to those used in real life (Arter and Spandel, 1992; Gifford and O'Connor, 1992). Among the assessment procedures that are currently gaining favor because of their realism and instructional relevance is portfolio assessment (Reckase, 1995). The issues that are often raised by portfolio assessment include portfolio definitions that are consistent with its purpose, a portfolio assessment goal to improve instruction or to evaluate learning, and data standards that include reliability and validity.

Many definitions of portfolios are available. Almost as many as there are portfolio projects. Three of those definitions will be reviewed here because they were the definitions that inspired the DISD portfolio assessment project. When portfolios are used to evaluate student performance, Barone (1991) defined them as collections of students' work over an extended period of time that are reviewed against criteria in order to judge an individual student or a program. The portfolio or collection of work does not

constitute the assessment; it is simply a receptacle for work (essays, videotapes, art, journal entries, and so on) that may or may not be evaluated. Another definition of portfolios stated, "A portfolio is a purposeful collection of student work that exhibits the student's efforts, progress, and achievements in one or more areas" (Paulson, Paulson, & Meyer, 1991, p. 60). The collection of work should not be confused with the assessment process. Teachers often keep portfolios of student work, but they are not portfolio assessments. Finally, intertwining the definition of a portfolio with its purpose Herman, Aschbacher, and Winters (1992) stated that the "assessment" in portfolio exists only when (1) an assessment purpose is defined; (2) criteria or methods for determining what is put into the portfolio, by whom, and when, are explicated; and (3) criteria for assessing either the collection or individual pieces of work are identified. These three definitions served as the basis for DISD's development of a portfolio definition.

When portfolio assessment first became a topic under the authentic assessment movement, it was clearly based upon a strong classroom setting with teachers and students developing the portfolio. Researchers were strongly emphasizing a bottom up approach to portfolio design at that time. Critical to the definition of a portfolio assessment was the role of the teacher. Portfolio implementation in the DISD followed the then current research and designed a bottom up portfolio beginning at the classroom level. Some of the research available then included Herbert (1992) who stated that the teacher's role in a portfolio assessment was defined by a two-part process. First, decisions must be made by the teacher about the collection process. What goes in, who chooses, when samples are taken--these are dimensions of the assessment task that define the setting and kinds of work that will be considered. Second, the scoring criteria must be defined by the teacher in collaboration with the student(s).

The DISD Chapter 1 program staff were looking for results reported by Aschbacher (1993) when he suggested that teacher involvement in the development of portfolios influenced teachers' instructional practices, the way they thought about their teaching, and their attitudes toward their students. As Koretz et al. (1993) stated, "Although the amount of change reported by most teachers was small, the pervasiveness of change was striking" (p.23). Portfolio design questions in the literature which greatly influenced the DISD portfolio design process were raised by Arter and Spandel (1992). In looking at the teacher involved design issues they stated that, "A grass-roots effort not only has the potential to improve

instruction, but also to produce the rich and valid sources of information needed for better large-scale assessment" (p. 39). They reported that portfolios mandated from on high are often seen as "impositions into both students' and teachers' time and for that reason the content was not likely to be valid" (p. 39). This was further demonstrated by the Vermont Portfolio Assessment Project. Koretz et al. (1994) reported that the Vermont portfolios were primarily focused on unstandardized tasks. Students and teachers had nearly unconstrained choice in selecting tasks to be placed into the portfolios. The program was truly a bottom up design with committees primarily responsible for developing the operational plans for the program. They were complemented by a single, standardized prompt scored with the same rubrics used with other portfolio items. Participation in the program demanded a lot of time and substantial resources; however, it had a very powerful effect upon instruction. Based upon this and other literature that was reporting multiple affects in instruction and assessment from a bottom up portfolio design, the DISD designed its portfolios similarly.

Recently, Shepard (1995) reported on a classroom-based performance assessment project begun two years ago. Teachers were heavily involved in the assessment design for the study. Some of the issues that Shepard found were problems included teacher time and familiarity with assessment issues. After working with teachers in the project for two years she concluded that "teachers eventually developed greater sophistication about scoring criteria...teachers were much more aware that scoring rules should depend on what you were scoring for (the intended construct, in measurement terms)" (p.41). This study was occurring concurrently with the DISD portfolio project. Similar problems were found. Using a bottom up approach takes lots of teacher time and on-going staff development activities. The results are shown in changes in classroom instruction and an increased knowledge for teachers of the assessment issue itself. If portfolio assessment has as a primary goal the improvement of classroom instruction, then it is appropriate to use a teacher-designed system in a bottom up approach. However, difficulties with the quality of aggregateable data needed for an accountability system were seen when a bottom up development was used. Unfortunately, unlike the Shepard study, the DISD portfolio project was also designed to report results for accountability purposes even during development.

Data reliability and validity issues came to the forefront when portfolio assessment was used in large-scale assessments. In examining the literature for reliability and validity issues the Vermont Portfolio Project must again be reviewed (Koretz et al., 1993, Koretz, et al., 1994). The purpose of the portfolio was to report statewide student achievement at the school level. This forced the issue of the reliability of the data because the results were being reported as school to school comparisons. The Vermont portfolios were analyzed at three levels. First, each piece in the portfolio received a score. Next, the scores were combined across common dimensions and finally the entire portfolio was scored by combining across all pieces and dimensions. In the Vermont program interrater reliabilities of .28 to .60 were obtained depending upon how the scores were aggregated. The low reliabilities were not sufficient to allow reporting many of the aggregate statistics Vermont had planned to use. Also it was found that there was participation by only a small number of students per school. Because of the low reliability of scores the issue of validity was somewhat moot. The implications of the Vermont study indicated the need to set realistic expectations for the program, to acknowledge the large costs, and to monitor the implementation and impact of the quality of the performance data they yielded. The authors recommended that portfolio tasks be standardized to improve the reliability of the scores and that raters receive further training to increase their level of accuracy. As the author warned, "Such standardization, however, runs contrary to many of the basic goals of portfolio" (p. 27).

Two other studies have also reported interrater reliabilities. In Pittsburgh, (LeMahieu et al., 1993) a large federally-funded portfolio assessment of writing portfolios had interrater correlations ranging from .60 to .70. In a separate, smaller study Herman and Winters (1993) reported average correlations for raters at .82 with the percentage of absolute agreements for all pairs of raters at .98. What was the difference among these reported reliabilities? When the contents of portfolios are relatively uniform and when experienced scorers use well-honed rubrics, the results show more interrater reliability. Thus the move to standardize and create portfolio assessments with greater reliability began.

Other studies that have addressed the data quality issues included Herman and Winters (1994) who concluded that technical quality is a critical issue if results are used to make important decisions about students, teachers, and schools. Portfolios need to provide accurate information for the decisions to

be made. Are the results reliable, consistent, and meaningful estimates of what students know and can do? They stated that one basic requisite for technical quality--interrater reliability--was achievable. "On the one hand, some of these technical issues can probably be most easily solved if portfolio tasks are closely specified and highly standardized. But, in seeking technical rigor, we need to be sure not to lose the appeal of the portfolio concept" (p. 54). While an admirable goal for portfolio assessment, it will be difficult to achieve if not impossible.

Validity issues have been very sparsely studied beyond claims that it "looks like" it captures important learning. Koretz et al. (1993) correlated the Vermont program results and direct writing assessment and found a range from .47 to .58 correlation. The same weak relationship was reported between a mathematics portfolio and a uniform test score in mathematics. Similarly, Gearhart (1993) found almost no relationship when comparing results from writing assessments and writing portfolios.

As a summary for this literature review, we quote from a recent study by Mark Reckase (1995) who addressed the issue of reliability by presenting one hypothetical model of writing. He found that a well-structured and carefully scored portfolio assessment had the potential to yield scores meeting the standards for reliability required for use with individual students. However, the cost was considerable. He suggested that portfolio assessment could be used in formative evaluation in the classroom emphasizing instruction and not require high levels of reliability. Such use has also been suggested by Moss et al. (1992). Other reliability issues including the stability of scores across time, across different rating groups, and the effect of task has not been reported in the literature on portfolio assessment.

Background Information about the Chapter 1 Program in the DISD

The DISD is one of the largest public school systems in the country. In the 1993-94 school year, the federally funded Chapter 1 compensatory education program in the DISD served a total number of 17,366 K-3 students in 98 elementary schools. This represented more than 40% of all K-3 students enrolled in a Chapter 1 school. The Chapter 1 program in Dallas focused on providing supplemental reading and language arts instruction to low achieving students. Chapter 1 schools in Dallas vary in sizes -- student enrollment in the schools ranged from 220 to more than 1,300. However, all shared two characteristics in common: (a) a high concentration of students from low income families and (b) a high

concentration of students from minorities families. At the District level, Hispanic and Black students made up the two largest ethnic groups enrolled in the program. White students only accounted for 6% of all Chapter 1 students.

Student assessment and program evaluations for Chapter 1 have historically been based upon norm-referenced and criterion-referenced tests in the areas of reading and language arts. The DISD utilizes the *Iowa Tests of Basic Skills (ITBS)* to obtain national comparison data and the *Texas Assessment of Academic Skills (TAAS)* to determine student mastery of the Texas Essential Elements.. The *TAAS* is a state-mandated multiple choice test administered in grades 3 and above to assess student's abilities to read and compute. The Chapter 1 program also uses these measures to determine the effectiveness of the Chapter 1 program and to decide whether or not schools are meeting the Chapter 1 established goals annually.

Initial Development of the Portfolio Assessment for the DISD Chapter 1 Program.

In the Spring of 1992, the Texas Education Agency's (TEA) Standard Application System (SAS) 201 for Chapter 1 funds required that Districts statewide include other outcome measures as well as standardized, norm-referenced tests in their applications. The SAS 201 stated that one of these other outcome measures could be a portfolio collection. The portfolio had to include two types of documents—a district-developed checklist and samples of the student's work. Districts were required to set a goal for the portfolio in terms of what percentage of students would "master the Texas Essential Elements" which were the mandated basis for the portfolio. Within those parameters the District began developing its first portfolio assessment for the 1992-93 school year.

Based upon the literature written at that time the DISD began with a bottom up approach to portfolio design where teachers tried out portfolio assessment in the context of classroom instruction. It was believed at that time that this would provide valid and reliable data and also improve classroom instruction as teachers became more knowledge about the issues included in a portfolio assessment. From the beginning, however, the conflicting goals of portfolio assessment struck us as program evaluators as being at odds.

Several decisions were required as the first portfolio assessment was designed and implemented within the District. First, what goal would be set Districtwide as the percentage of students mastering the Essential Elements as measured by the portfolio? Without historical data to base a decision upon, it was impossible to know what would be a reasonable outcome. Further complicating that decision was the knowledge that the Districtwide goal had to be applied to all campuses within the District.

The goal was finally determined from the historical performance of students on the *TAAS* objectives. Since the *TAAS* purportedly measures some of the Texas Essential Elements, it seemed reasonable that students would perform on a portfolio assessment at least as well as students had been performing on the *TAAS*. Therefore, the District's goal was set at 40% mastery. A 40% goal required numerous explanations throughout the school year, but was accepted as reasonable once explained. We, of course, were hopeful that students would perform better on a portfolio assessment than they had been performing on a traditional multiple-choice test.

Next, the District needed to define what was mastery? What would the District-developed checklist contain and how would it be structured? How would the portfolio be rated and who would do the rating? How would the results be reported and what timeline would be used for recording, rating, and reporting? Would the portfolios be monitored by staff outside the classroom or campus? What samples of the student's work would be included and how many samples did there need to be in each portfolio? The first year of portfolio assessment was spent trying to answer the myriad of questions being raised.

Once the SAS 201 was approved by the TEA, work began on the contents of the portfolio. Keeping in line with our belief that a bottom up design would yield not only valid and reliable data, but positively impact classroom instruction, a committee of teachers worked with the Chapter 1 instructional specialists for three months to define mastery, develop the District's checklist, and determine portfolio contents. As a result of that committee's work the first Chapter 1 Essential Elements Checklists were published in December 1992. It was decided that Chapter 1 teachers would score the results of their own students' portfolios and report those results to the evaluation staff in the form of a three-point mastery scale which would be called *Some of the Time*, *Most of the Time*, and *Not Yet*. Since all of the Texas Essential Elements were to be evaluated in each portfolio, the rubric varied by grade level as the number

of Essential Elements (11 to 16) changed across grades. Teachers initiated the portfolio assessments in January 1993.

The committee of teachers who defined the three-point mastery levels as Some of the Time, Most of the Time, and Not Yet left the interpretation of what each of the three levels meant to each teacher. They believed that structuring the portfolio with a cut-off score for the three levels would be too restrictive and that teachers needed to use their professional judgment. Since all portfolios would be different collections of student work and the District developed checklists were also open to interpretation by the teacher, the committee reasoned that the mastery status for each Essential Element could not reasonably be District defined.

The problems reported at the end of the first year of portfolio assessment implementation covered four areas. First, there was not full districtwide implementation of the portfolio assessment process. A few schools did not keep portfolios, some kept all of the student's work, others kept no student work samples. Guidelines were needed for the collection process as well as the implementation that was required. Second, a concise definition of mastery was needed for teachers to use in evaluating their portfolios. Next, the Chapter 1 program needed to focus the scope of the portfolio to a limited number of Essential Elements that were measurable. Finally the quality of the student work samples needed to be addressed. Too many of the work samples were simply workbook pages completed by the student and not clearly related to the Essential Elements.

Improvement of Portfolio Assessment in the DISD in the Second Year

As in the 1992-93 school year, portfolio assessment was required in 1993-94 by the TEA as one of the other outcome measures used to evaluate the Chapter 1 program. Therefore, the DISD Chapter 1 program, for a second time, adopted portfolio assessment to measure Chapter 1 students' mastery of the Texas Essential Elements and to evaluate, along with other measures, the effectiveness of the Chapter 1 program in the 98 Chapter 1 schools. The District's goal was that 48% of the Chapter 1 students would master the Texas Essential Elements.

The second year of Chapter 1 portfolio assessment was a continuation of the breakthrough work of the first year. Following the rules laid out in 1992-93, Chapter 1 teachers were required to develop a

portfolio folder for each Chapter 1 student. The portfolio had to contain a Texas Essential Elements checklist (as the one used in the first year) and a minimum of six selected student work samples. Chapter 1 teachers were asked to observe each student's performance during the period from December 1993 to April 1994 and make judgments about the student's mastery level on a three category scale (Not Yet, Some of the Time, and Most of the Time). The work samples were required to be representative of a student's work and be referenced to the items on the checklist to serve as evidence supporting the teacher's judgments.

The second year of Chapter 1 portfolio assessment was a development upon the first year. Based upon our experiences from the first year we realized that there were several key issues that needed to be resolved. Efforts were made in the following aspects to improve the implementation. However, we did not realize then that most of the problems were related to the bottom up design of the District's portfolio assessment. If the data had been used only for classroom instruction then these issues would not have been considered problems. They became problems because of the uses made of the District's portfolio assessment. Because of our failure to realize this at the time, we only embarked upon methods to improve the current process and did not realize the greater implications.

1. Ensure complete implementation of *the portfolio assessment as a large-scale evaluation tool*

Change theories report that whenever programs are revised the first step in implementing the change is to overcome the resistance to change. This was seen in the first year of portfolio assessment. Some schools were not complying with the Chapter 1 requirements (did not have a portfolio for every Chapter 1 student) and that some of those that were complying were not collecting student work samples or were not using the District-developed checklists. Two steps were taken to solve this problem.

The first step was the development of guidelines for implementation. These included detailed requirements for portfolios and data collection procedures. More importantly, we found the most practical and effective way to ensure implementation was to send monitors to each of the schools to review and report on the contents of each portfolio. The first year monitors were sent to each of the Chapter 1 schools in April to review the contents of the portfolios. The monitors did not check beyond describing the types of student work (work sheets, audio tapes, writing samples) in the portfolio, counting the number of

documents in the portfolio, and looking at the checklist notations made at the time. In the second year, Chapter 1 monitors visited schools in February and May 1994. The monitors counted the total number of portfolios maintained at schools by grade and reviewed the contents of one-third of the portfolios to see if a proper portfolio was developed for every Chapter 1 student. The February monitoring results were summarized and reported to schools before the May monitoring. The final results show that at the District level 99% of Chapter 1 students had a portfolio maintained by their teacher, 82% of the reviewed portfolios contained a completed District-developed Essential Elements checklist, and 65% of them carried seven or more pieces of student work.

2. Make the criteria more meaningful and demonstrable to teachers

First, in the process of implementation we realized that the inclusion of all of the Texas Essential Elements were too large for teachers to work with under a portfolio assessment. Since the Texas Essential Elements included anywhere from 11 to 16 Essential Elements depending upon grade level and English or English-as-a-second language. Many of the Essential Elements were not measurable or documentable either. During monitoring it was found that many of the portfolio collections of student's work only focused on a few of the Essential Elements. As a result a Teacher Handbook was created for the current school year focusing upon a few of the Essential Elements across all grades and guiding teachers in ways to collect and evaluate student work in these Essential Elements.

Second, the use of vague mastery categories limited the interpretation of individual teachers as they tried to use them in grading a student's portfolio. What is mastery? What is not? The teacher committee had insisted that each essential element be scored on three levels and that the teachers would not determine overall mastery or non-mastery. Because of the need to aggregate student performance across grades, the program evaluator set a mastery, non-mastery definition after reviewing the results of the first-year portfolios. Mastery was defined for each student as scoring Most of the Time or Some of the Time and not receiving a score of Not Yet on any of the Essential Elements. Even though this was an improvement it was still not clear enough for teachers to be able to resolve scoring difficulties. This problem remains unresolved to this date. The lack of consistent understanding and interpretation of the

judgment scales among teachers no doubt affected the reliability of the portfolio assessment. Even though it was unable to be dealt with in 1993-94, this issue is critical.

3. Maintain quality in collecting student work samples

Another improvement effort made in 1993-94 was related to the quality of student work samples contained in the portfolios. Several quality problems with Chapter 1 portfolios were found during the first monitoring, such as the lack of link between the Texas Essential Elements (criteria) and student work samples; and the inappropriate type, number, and quality of work samples. These problems led to a large number of Chapter 1 portfolios that failed to provide meaningful information about the level of student performance. To solve the problem, a memo addressing these issues along with specific rules for selecting work samples were sent to schools. As a result, progress was made by the end of the year as noted by changes between the first and the second monitoring visits.

4. Quantify data and aggregate individual student scores across grade and school

In 1993-94 an effort was also made to quantify individual data into a comprehensive score. After Essential Element Mastery forms were returned from each campus, the three categorical scales were quantified into numerical scales (Not Yet = 0, Some of the Time = 7, Most of the Time = 10). Scores of each element on the mastery form were combined to form a single total score indicating each student's overall mastery level. A student was considered to have attained mastery if his/her score was equal to or above 70% of the highest possible score. The highest possible score was the number of items on the Texas Essential Elements list times 10 (the largest point value for one item). For instance, if there were 13 items on the list, the highest possible score was 130 points. An individual school's mastery rate, or the percentage of students meeting the mastery level at each grade, was then calculated. Schools with a mastery rate of 48% or above were categorized as successful. The overall District mastery rate was also calculated. The 1993-94 final evaluation results indicate that 71% of the Chapter 1 students Districtwide passed the mastery level and 88 of the 98 Chapter 1 schools met the District goal by achieving 48% or above mastery.

Results of Validity Tests on the DISD Chapter 1 Portfolio Assessment

As a performance-based evaluation approach, portfolio assessment was in its second year of implementation in the DISD. It was important to know if portfolio assessment could be used as a substitute for a traditional paper and pencil, multiple-choice measure of the same or very similar objectives. That is, does portfolio assessment have concurrent validity? To examine the concurrent validity, the relationship between third-grade Chapter 1 students' performance on Essential Elements Mastery Forms and their scores on the *TAAS* Reading test were examined at the end of the 1992-93 school year and again at the end of the 1993-94 school year. It was assumed that if portfolio assessment is a valid alternative measure of the Texas Essential Elements, the results of the portfolio assessment should, to some extent, correlate with the results of the *TAAS*.

The results of comparing the third-grade Chapter 1 results in the 1992-93 school year for the *TAAS* in September and the portfolio assessment in May was not encouraging. While 4% of the Chapter 1 third graders tested with the *TAAS* met minimum expectations and only 1% mastered all objectives, based upon the portfolio assessments by the end of third grade, 43% mastered the Essential Elements Most of the Time.

The concurrent validity of Chapter 1 portfolio assessment was again examined at the end of the 1993-94 school year. Two third-grade Essential Elements were selected because they closely matched two third-grade *TAAS* reading objectives. The first, "develop vocabulary to understand written language in meaningful context," was matched with the *TAAS* objective "word meaning." The second, "use comprehension strategies to construct meaning from the text," was matched with the overall *TAAS* reading test which included five reading comprehension objectives. For each pair of objectives, teacher portfolio ratings were compared with *TAAS* mastery status. The data were aggregated at the school level first, and then at the District level. The results of the comparisons are presented in Tables 1 and 2.

Table 1 displays data from a comparison between *TAAS* word meaning mastery status and teachers' portfolio vocabulary development ratings for Chapter 1 Grade 3 students. Overall, the results indicate a positive, but very weak, association between the two measures. Consider two observations. First, the overall mastery rate for *TAAS* (54%) was lower than that of portfolio ratings (49% - Most of the

Time and 45% - Some of the Time). Yet, despite the differences in the mastery rates of the two measures, the students who mastered the *TAAS* were more likely to have been rated as masters on the portfolios (59% of *TAAS* masters were rated mastery - Most of the Time) than students who did not master the *TAAS* (39% of *TAAS* non-masters were rated mastery - Most of the Time). Conversely, a smaller percentage of *TAAS* masters than non-masters were rated as non-masters on the portfolios (3% versus 9%).

Table 1

Crosstabulations of *TAAS* Word Meaning Mastery Status and Teacher Portfolio Vocabulary Development Ratings for Chapter 1 Grade 3 Non-LEP Students

	TAAS Mastery Status					
	Mastery		Non-Mastery		Total	
Portfolio Ratings	N	%	N	%	N	%
Mastery - Most of the Time	740	59	407	38	1,147	49
Mastery - Some of the Time	479	38	576	53	1,055	45
Non-Mastery - Not Yet	38	3	99	9	137	6
Total	1,257	100	1,082	100	2,339	100

Table 2 presents data from the comparison between *TAAS* reading comprehension mastery status and teacher portfolio ratings of student use of comprehension strategies. Similar results were found from this comparison. Despite the overall higher mastery rate of the portfolio rating (44% - Most of the Time) than the mastery rate of the *TAAS* mastery (33%), the students in the *TAAS* mastery were more likely to have been rated as mastery - Most of the Time (60%) than the students in the group that did not master the *TAAS* (36% were rated mastery - Most of the Time). Again, a smaller percentage of *TAAS* masters than non-masters were rated non-masters on the portfolios (2% versus 11%).

Table 2

Crosstabulations of *TAAS* Reading Comprehension Mastery Status and Teacher Portfolio Ratings of Student Use of Comprehension Strategies for Chapter 1 Grade 3 Non-LEP Students

	TAAS Mastery Status					
	Mastery		Non-Mastery		Total	
Portfolio Ratings	N	%	N	%	N	%
Mastery - Most of the Time	464	60	563	36	1,027	44
Mastery - Some of the Time	295	38	835	53	1,130	48
Non-Mastery - Not Yet	12	2	170	11	182	8
Total	771	100	1,568	100	2,339	100

Overall, these results indicate that portfolio assessment has certain "skewed validity" in measuring the Texas Essential Elements. More *TAAS* masters than non-masters were rated as masters on the portfolios. Fewer *TAAS* masters than non-masters were given non-mastery ratings on the portfolios. Yet, it was much more difficult for a student to pass the *TAAS*. *TAAS* non-mastery rates were much higher than portfolio non-mastery rates. The higher passing rate was, we believe, due in part to a lack of understanding about the criteria to be used in scoring the portfolios.

Discussion of Unsolved Problems from the Second Year

Despite the great effort made in 1993-94 to improve the quality of the portfolios, due to the nature of the initial design of the District's portfolios many problems remained unsolved. A key weakness found in 1993-94 portfolio assessment is the lack of a link between the criteria and students' work samples. Consequently, work samples contained in the portfolio did not provide meaningful information to support teachers' judgment about students' performance levels. In order for teachers to select appropriate work samples, criterion behaviors on the checklist need to be task-oriented, specific, and demonstrable. The current criteria, the Texas Essential Elements, did not meet this standard. It was recommended that the criterion behavior be refined so that they are more manageable for teachers. It was also recommended that teachers select appropriate work samples matching the criteria. A rubric needed to be established to define the type, number and quality of the work samples. For example, worksheets and art work may not be a valid indicator of a student's mastery level and therefore are not desirable work samples.

The nature of portfolio assessment leads to a large variation (or individual style) in the implementation. The lack of consistent understanding and interpretation of the judgment scales among teachers no doubt affected the reliability of the portfolio assessment. This issue must be addressed if any measure of reliability is to be applied to the assessment system. In the Dallas model teachers are individually scoring their own students' work without the benefit of a second rater. Therefore, a measure of interrater reliability is not available to the District. The current portfolio assessment program has not been redesigned to correct this major reliability problem. A second-rater system was recommended at the end of the 1993-94 school year.

Another portfolio design problem raised in the 1993-94 practice concerns the assessment "timeline." A portfolio should contain work samples that illustrate student's growth. The 1993-94 Chapter 1 portfolio assessment was not designed to indicate the students' mastery level change from the beginning of the year to the end of the year. It was recommended that a timeline be included along with the checklist and rating scale.

The dichotomy of having a grass-roots measure of student performance which is tied to real life experiences used to document program accountability remains unsolved at this time. It is difficult to take a measure that uses vague definitions of mastery levels and interpret that meaningfully. The very nature of portfolio assessment leads to a large variation (or individual style) in its implementation. However, such variation and the lack of consistent understanding and interpretation of the mastery scales among teachers no doubt affects the reliability of the portfolio assessment. More clearly defined, narrowed-down scales was among the strategies recommended at the end of the 1993-94 school year. But this recommendation flies in the face of the original meaning of portfolio assessment as adopted by the District. Even this year as a districtwide scoring rubric was mandated from the top down to the classroom there are still clarity issues.

Implications for Portfolio Related Policies

During the last two years we were enmeshed in the DISD's design, implementation, monitoring, and reporting of a portfolio assessment system. We were vaguely aware that there was great conflict within the process of portfolio assessment, but we were so involved in the minutia of the process that we could not see the overarching problem. Now that we have had time to reflect upon the issues, we can clearly see that there is a need to redefine the concept of portfolio assessment.

Because the purpose of any assessment must dictate the type of assessment to be used, then it only seems logical to conclude that portfolio assessments are truly of two types each with a different purpose as shown in Figure 1. When a portfolio is used for accountability purposes then it must be designed from the top down with clearly defined criteria and appropriate rubrics. The freedom of allowing individual teachers and students to change it must be restricted. Accountability requires reliability. Reliability in portfolio assessment comes only from a well-structured and carefully scored portfolio. When a portfolio is

used to improve instruction and learning, then it must be designed from the bottom up allowing the individual teachers and students freedom to create a portfolio that is unique. Staff development for the teachers must be part of this process so that issues such as scoring criteria are included.

Current problems that policy makers need to understand are that they are mandating that Districts use a top down approach to reach bottom up goals. These are in conflict and are creating a great deal of confusion about appropriate uses of portfolio assessment nationwide. The prior literature (Arter and Spandel, 1992) was wrong. It is, in our opinion, not possible to use a grass-roots effort to improve instruction and to also produce valid sources of information needed for better large-scale assessment. Grass-roots development of a portfolio assessment will improve instruction, but will not serve the purposes of large-scale assessment.

Before you implement portfolio assessment we strongly recommend that the main purpose of a portfolio assessment be clearly defined. If you want to use it as an accountability tool then first you must have a standard, the contents need to be clearly identified, the scoring criteria need to be specified in detail, and all who work with the portfolio need to understand the nature of the portfolio. If you want to use a portfolio assessment to improve classroom instruction then a bottom up approach is most appropriate. The parameters can be less clearly defined so that teachers will have the freedom to work within the nature of a portfolio to refine it as needed and as they improve and change instructional methodologies. You cannot utilize a single portfolio assessment system to accomplish both goals.

We would argue that too much time has been spent trying to fit a complex reality into one theory. Rather than blending the insights of different portfolio assessment purposes into one explanation that impoverishes the quality of all interpretations, it is time to recognize that there are truly two types of portfolio assessments--an accountability model and a classroom-based model.

Figure 1

Two-models of Portfolio Assessment

Instructional Improvement Model	Accountability Model
Purpose: Improve classroom instruction	Purpose: Provide valid and reliable data to determine program effectiveness at the school level
High teacher involvement	Low teacher involvement
Teacher selected portfolio contents	Centralized content selection
Mastery objectives individualized to student/classroom needs	District or schoolwide objectives determined by the program
Teachers use their professional judgment to determine the mastery level of each student	Use specific rubrics districtwide to determine the mastery level of each student
Teacher/student decide what to put into the portfolio	Portfolio contents are mandated.
Scoring done by individual teacher(s)	Use scoring rules or second raters to improve scoring consistency across teachers
Monitoring is not needed	Process monitoring is important
Score summarization and school level aggregation are not needed	Score summarization and school level aggregation are needed
Potential for low level of technical quality (reliability and validity)	Potential for high level of technical quality (reliability and validity)

Bibliography

Arter, J. A. & Spandel, V. (1992). Using portfolios of student work in instruction and assessment. *Educational Measurement: Issues and Practice, 11(1),* 36-43.

Aschbacher, P. R. (1993). *Issues in Innovative Assessment for Classroom Practice: Barriers and Facilitators.* (Tech. Rep. No. 359). Los Angeles: University of California, CRESST; Center for the Study of Evaluation.

Barone, T. (1991). Assessment as theater: staging an exposition. *Educational Leadership 48(5),* 57-59.

Final Evaluation Report of the 1992-93 Chapter 1 Instructional Program. (1993). Dallas, TX: Dallas Independent School District.

Final Evaluation Report of the 1993-94 Chapter 1 Instructional Program. (1994). Dallas, TX: Dallas Independent School District.

Gearhart, M. & Herman, J. L. (1995). Portfolio assessment: whose work is it? *Evaluation Comment.* Los Angeles: UCLA's Center for the Student of Evaluation & the National Center for Research on Evaluation, Standards, and Student Testing.

Gearhart, M., Herman, J. L., Baker, E. L. & Whittaker, A. (1993). *Whose Work Is It? A Question for the Validity of Large-Scale Portfolio Assessment* (Tech. Rep. No. 363). Los Angeles: University of California, CREST: Center for the Study of Evaluation.

Gifford, B. R., & O'Connor, M. C. (Eds.). (1992). *Changing Assessments: Alternative Views of Aptitude, Achievement and Instruction.* Boston: Kluwer.

Herbert, E. A. (1992). Portfolios invite reflection from students and staff. *Educational Leadership 49(8),* 58-61.

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A Practical Guide to Alternative Assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

Herman, J. L. & Winters, L. (1995). Portfolio research: A slim collection. *Educational Leadership, 52(2),* 48-55.

Koretz, D., Klein, S., McCaffrey, D. & Stecher, B. (1993). *Interim Report: The Reliability of the Vermont Portfolio Scores in the 1992 School Year* (CSE Tech. Rep. 370). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Koretz, D., Stecher, B. & Deibert, E. (1993). *The Reliability of Scores from the 1992 Vermont Portfolio Assessment* (Tech. Rep. No. 355). Los Angeles: University of California, CRESST: Center for the Study of Evaluation.

Koretz, D., Stecher, B., Klein, S., McCaffrey, D. (1994). *The Vermont Portfolio Assessment Program: Findings and Implications.* Washington, D.C.: U.S. Department of Education, Office of Educational Research and Improvement.

LeMahieu, P. G., Gitomer, D. H., & Eresh, J. T. (1993). Portfolios in large-scale assessment: Difficult but not impossible. *Educational Measurement: Issues and Practice. 12(3),* 12-21.

Moss, P. A., Beck, J. S., Ebbs, C., Matson, B., Muchmore, J., Steele, D., & Taylor, C. (1992). Portfolios, accountability, and an interpretive approach to validity. *Educational Measurement: Issues and Practice, 11(3),* 12-21.

Paulson, R. L., Paulson, P. R., & Meyer, C. (1991). What makes a portfolio a portfolio? *Educational Leadership 48(5),* 60-63.

Reckase, M.D. (1995). Portfolio assessment: a theoretical estimate of score reliability. *Educational Measurement: Issues and Practice 14(1),* 12-14.

Shepard, L. A. (1995). Using assessment to improve learning. *Educational Leadership 52(5),* 38-43.